DOCUMENT RESUME

ED 122 613 FL 007 601

AUTHOR Goodrich, Hubbard C.
TITLE Distractor Efficiency: A Study into the Nature of Distractor Efficiency in Foreign Language Testing.
PUB DATE Aug 75
NOTE 34p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Arabic; *English (Second Language); *Item Analysis; Language Instruction; Language Proficiency; Language Skills; *Language Tests; *Multiple Choice Tests; Native Speakers; Second Language Learning; *Test Construction; *Vocabulary
IDENTIFIERS *Distractor Efficiency (Testing)

ABSTRACT

The aim of the research in question was to investigate the efficiency of various classes of distractors used in multiple-choice vocabulary question testing. Much of the quality of multiple-choice questions relies on the extent to which incorrect choices tempt the less proficient student. The study selected 8 classes of distractors and attempted to discover whether their effectiveness could be quantified and a rank order of preference could be established for population samples of different levels of English proficiency. The study, involving 1200 Arabic-speaking students of English, found a consistent order of preference and some puzzling questions. It was found that a definite hierarchy of distractor efficiency exists in terms of discrimination and, particularly, potency. All groups demonstrated a preference for distractor types, but there was an uneven agreement in discrimination. In addition, the most popular distractors were not necessarily the most discriminating. Differences between populations of varied proficiency were greater than between groups from different geographical locations. (Author/KM)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED122613

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

A STUDY INTO THE NATURE OF DISTRACTOR EFFICIENCY IN FOREIGN LANGUAGE TESTING

by

H. C. GOODRICH

University of Petroleum and Minerals
Dhahran - Saudi Arabia

- FL007601

DISTRACTOR EFFICIENCY

The Multiple Choice Test format has been selected by many language teachers for the measurement of student achievement or general proficiency. Writing such a test, however, is not easy. The professionally developed test utilises procedures unavailable to the classroom teacher. The latter usually has little training, equipment or time in the construction of a class test to exercise some of the arcane operations of the testing specialist. Yet most teachers are required to prepare class tests at some point in their teaching careers. How can the teacher improve his test writing efforts? Let us first look at the type of test questions being considered.

The structure of the multiple choice question under scrutiny here has three basic elements. (1) the Lead, Stem, or Question (2) the Correct Choice or Answer, and (3) two to four Wrong Choices or Distractors.

Lead	:	(1) +	I write with a_____ .	
Answer	:	(2) +	(a)	pen
Distractors	:	(3) +	(b)	horse
			(c)	book
			(d)	fork

The effectiveness of Multiple Choice Questions depends on the validity of the lead and the correct choice, but we must not lose sight of the importance of the distractors in making any question of this type efficient. Choosing distractors is a small but essential aspect of constructing multiple choice items. Much has been written about tests and measurement, but few studies seem to have been undertaken to investigate the area of alternate choices. Specialists in foreign language testing have mentioned several types of distractors[1] but it has been difficult to find evidence to substantiate their selections.

An effective question and each distractor should have a degree of potency and discrimination.[2] The degree of potency is measured by the number or percentage of students making a specific choice. If no one chooses a particular distractor, it is not participating effectively in the process of giving the question a factor of difficulty. On the other hand, a distractor which is too attractive may be a correct answer to a badly posed question.

A distractor's facility of discrimination is its ability to separate or differentiate students of differing levels of proficiency - between able and less able students. The potency of a distractor must be combined with its factor of discrimination if its efficiency is to be determined. One quality without the other can be misleading.

Most of us who have written tests have had the experience of

writing what we think are good leads but discover that all the students select the correct answer. No one has been fooled by the alternate choices. Tests should place students into levels of achievement or proficiency. If our questions are crippled, we have failed to provide a realistic assessment of relative ability. Is it possible to improve the question in the initial selection of distractors?

A research effort was initiated in the summer of 1973 to determine whether classes of distractors could be ranked in terms of efficiency. It did not propose to study the type or content of the test questions; nor did it plan to delve into other aspects of the testing trade. It was designed to focus its attention exclusively on the distractor.

The purpose of this paper is to report the results of the investigation into the efficiency of various types or classes of single-word distractors in testing English as a Second Language in an attempt to improve a teacher's test writing capabilities in at least this one area and to answer three questions:

1. Which distractor category has the greatest potency to attract students away from the correct response?

2. Which distractor category has the greatest facility or power to discriminate or separate students at different levels of proficiency? and

3. What differences are there, if any, between populations

of the same language but from different geographical areas?

Our aim, then, is to identify and define important classes of distractors and to design a test situation that would force students to demonstrate the relative efficiency of a class of distractors rather than any one specific distractor as one might find in a normal item analysis. With the assistance from the Data Processing Center[3], a test format was designed. It was suggested that vocabulary items should be selected for simplicity of analysis, that completion leads be used, that two forms of identical questions but different distractors be developed, and that the final product be presented to as large a population as possible.

The first task was to identify discrete types of distractors. A search of the available literature and an exercise in imagination discovered ten distinct types which could be classified. Eight were finally chosen for the study.

1. False Cognate

A Cognate is a word similar in meaning and pronunciation in language one (L 1) and Language two (L 2). A false Cognate is a word that sounds like a word in L1 but does not mean the same thing in L2. An example might be tea in English and /ṣay/ in Arabic. The English word shy would be a false cognate. If one were to write a test item where the correct answer would be tea, an inept student with some familiarity with the English orthographic system might select

a distractor such as shy because it sounds to his inner ear like a choice that might be an appropriate response to the question.

2. Cloze

The cloze procedure involves the deletion of one or more words from a passage and asking the student to complete the passage with an appropriate word. The cloze distractor is an inappropriate word chosen by students of L2 to complete a cloze procedure test.

3. Antonym

An antonym is a word which has an opposite or near opposite meaning to the correct choice.

4. False Synonym

A false synonym is a word which has a close or, in some situations, a similar meaning to the correct choice, but a selection which is not appropriate given the context of the lead.

5. Graphemic Variation

A graphemic variation is a word with letter modification - a word with letters moved, parts changed or shifted in certain ways as in reversals, substitutions, or positional alterations which may or may not reflect phonemic contrasts between L1 and L2. Examples might be bread/beard or hide/ripe.

6. Arbitrary

An arbitrary choice is a word selected at random and is usually completely irrelevant to the lead.

7. Affix

An Affix choice is a word which is modified by morphemic addition or deletion.

8. Contextual

A Contextual choice is a word which is related to the context of the question. It is its topicality that might mislead a student.

A number of other distractors were rejected. A pragmatic distractor is a word selected on the basis of classroom experience and might be difficult to generalize and apply outside the teacher's personal experience. Misspellings and nonsense words are also possible categories of distractors but were not included in this study for reasons of personal prejudice.

The second task was the actual process of selection. As the false cognates were considered to be the most restrictive and, with cloze, the most time consuming to develop, the search for distractors was initiated with the first two types.

The false cognates were chosen by a multiple process of elimination. First, the author read all the stimulus words from a frequency list* of 3000 words (the syntactic or function words were rejected) to three bilingual Saudi informants. The informants would translate orally. The author selected those Arabic words which might sound like an English word. Three hundred and fifty-two words were discovered and the process reversed. The false cognate words were then printed on a list in English and shown to 50 Saudi students of English. They were asked to write down any Arabic word or words that were brought to mind for each English cue word. If 20% or more of the students agreed that the cue word called to mind a specific Arabic word with the same meaning as the original stimulus

*Thorndike - Lorge

word, it was selected as a false cognate. The first step involved 3000 words. The second step involved 352 words. Only 143 words remained after the third step.

The 143 words were then submitted to a cloze procedure test. Ten native speakers of English (British and American teachers) and twenty Saudi students of English were given the cloze test. The results were tabulated and compared. If two or more students selected an alternate response to the one chosen by the native speakers, it was retained as a distractor. If there was agreement between the Saudi students and native speakers, the word was discarded. This procedure eliminated all but 60 of the original stimulus words.

The selection of the first two distractor groups involved the most amount of time and energy. In fact, so much effort was expended that these two classes of distractors might be eliminated from consideration by teachers interested in developing their own objective class tests. Most teachers do not have the time to devote to such activities. However, all the other distractors were chosen on the basis of the 60 remaining stimulus words each with a cloze and false cognate distractor.

Antonyms and false synonyms were chosen from various dictionaries and a Roget's Thesaurus. Where a clear antonym or synonym did not exist, or where it involved obvious differences in levels of difficulty, alternates were selected on the basis of prima facie differences or similarities.

Graphemic variation distractors were selected by the manipulation

of the consonant letters of the stimulus word in an attempt to find some real English word. It was assumed that the similarity in appearance of the distractor to the stimulus word in a visual sense might be attractive to some students. If more than one alternative existed, the final selection was subjective. No purely spelling differences were chosen (e.g. dessert vs. desert).

Arbitrary choice distractors were chosen by chance from a 10,000 word frequency list. Any choice was eliminated if the number of syllables (word length) varied by more than two syllables, if the choice was a possible correct answer, or for syntactic incompatability.

Affix distractors were selected by the modification of the stimulus word using appropriate affixes. The number of these distractors was limited in this study as their use might obviously provide a student with a clue to the correct answer.

Contextual distractors were chosen on the basis of the semantic environment of the question. The author made the selection solely by reference to the lead.

The third step was to write the questions. Each question was a completion type lead which required an understanding of the item in order to be able to select the correct answer. Most of the questions retained vocabulary below the 3000 words level. It was decided to increase the number of questions on the test from sixty to eighty even though eight distractors were not available in all cases for all questions. The additional questions were designed

to increase the quantity of some categories with few examples and to add questions with proven difficulty and discriminatory capability to increase the spread of the test population*.

Two forms of the exam were written.** Each form had identical questions but in a different order (Number 1 on Form "A" was Number 51 on Form "B"). Most questions on the different forms had different distractors so that all eight distractors, if available, would be given an opportunity to attract similar populations under similar conditions. The correct choices and distractors were distributed equally among the two forms. Where eight distractors were not available, one or more categories were used again on the different form. The answer key looks like the following. The numbers refer to distractor class:

TABLE # 1

Form "A"

Q #	Choice				
	A	B	C	D	F
1	2	4	x	3	1
2	x	7	4	1	2
3	5	3	1	2	x

Form "B"

Q #	Choice				
	A	B	C	D	E
51	5	6	7	8	x
52	2	5	6	x	8
53	4	x	6	8	7

Two examples of the procedure may be found on the following page (Table #2).

A computer program was designed to organize all student responses by distractor class and to provide statistics to quantify the resulting data in uniform and standard ways.

---

* A summary of the percentage of distractors in each category may be found on Page # 12 as Table #3 (pe).

** Directions written in Arabic.

TABLE NO. 2 - SAMPLE SELECTION PROCEDURE

STIMULUS word	inspect	throw
Oral English	inspect	throw
Oral Arabic	/fa:iš/	/yarmy/
1. False Cognate Student Response	fattish 80 % inspect	army 95% throw
2. Cloze : Questions	Our supervisor must____every school he visits	How far can yo____ a ball?
L1 L2	inspect (90 %) remember(30 %); advice (20%)	throw (95 %); hit (5 %) shoot (20 %); rock(10 %)
3. Antonym	speak	catch
4. Synony, false	sight	go
5. Graphemic Variation	insect	worth
6. Arbitrary Choice	fortunate	count
7. Morpho-syntactic	---	throwing
8. Contextual	buy	clean
Test Question	These new radios may not work. I must____them before we sell them.	Please pick up that paper and_____it into the basket.

Form A, #34	Form B, #4	Form A, # 23	Form B, # 73
a. inspect	a. advice	a. throw	a. rock
b. sight	b. speak	b. army	b. worth
c. insect	c. inspect	c. rock	c. count
d. fortunate	d. buy	d. go	d. throw
e. remember	e. fattish	e. catch	e. clean

A practice administration of the tests was conducted in the early Spring of 1974 with a population of 334 Arabic speaking Saudi students of English at UPM. Corrections were then made on the examination, and the computer program was refined.

During the late Spring, Summer and Fall of 1974. 101 students were tested at the American University of Cairo (AUC). Forty Saudi secondary students, 25 Saudi adults and 728 additional UPM students were also tested. We were frustrated in our efforts to expand the test population to other student groups in the Kingdom and to Arabic speaking groups in adjacent countries. A total of 1203 Arabic speaking students of English eventually made up our test population.*

The relative efficiency of a distractor is demonstrated by two criteria - a factor of potency and a factor of discrimination.

POTENCY: The Potency Factor was determined by measuring the difference between expected responses and actual responses for each distractor group. The expected responses were determined by dividing the total number of times a distractor appeared on the test by the total number of incorrect choices on the exam.

This may be expressed by the following Formula:

$$PF = Pa - Pe$$

Where: PF is the Distractor Potency Factor

Pa: is the actual Percentage of Student Errors represented by the Distractors. A figure which fluctuates depending on the group taking the test.

Pe: is the Expected Percentage based on the frequency of the Distractor's appearance on the exam. A fixed figure for all groups taking the test.

---

* A description of each group may be found in the Glossary on Page # 25.

A positive Potency Factor indicates the Distractor was chosen by students more often than would be expected according to the number of times the Distractor appeared on the test. A negative PF value indicates it was chosen less often than would be expected. The Formula above was used to provide the figures to rank the Distractor types according to their relative potency.

TABLE # 3 - POTENCY RANK FOR TOTAL POPULATION*

Distractor Class	Pa	Pe	PF	Rank
Cognate False	10.0 %	14.8 %	- 4.8	8
Cloze	9.3 %	9.3 %	0.0	4
Antonym	14.6 %	12.3 %	+ 2.3	3
Synonym False	18.6 %	12.1 %	+ 6.5	1
Grapheme	11.9 %	14.6 %	- 2.7	6
Arbitrary	9.1 %	13.8 %	- 4.7	7
Affix	7.7 %	8.7 %	- 1.0	5
Context	18.5 %	14.0 %	+ 4.5	2

* Figure and Ranking for each sub group may be found on Page #26, Table #6.

DISCRIMINATION: The Discrimination Factor was determined by the relationship between a student's total score on the test and the number of times he selected a given

class of distractor. Two points should be pointed out from the first:

1. The validity of the test may be rightfully questioned. No attempt has been made to provide construct validity for placing students in high or low groups. However, limited data did exist for some students and it seems safe to assume that a limited content validity existed.

2. The derived relationships are spurious because the student's score is a negative function of the total number of incorrect responses on the same test; therefore, the relative discrimination of the distractor will be determined by ranking the distractor according to the magnitude of the statistic measuring discrimination and not by the significance level of the statistic.

The following three statistical procedures were used initially to determine the ability of the distractor to differentiate between students with high and low scores:

i. The Pearson Correlation Coefficient is based on the linear relationship between the number of times a student chose a particular type of distractor and his score on the test. The discrimination power of the distractor would be inversely proportional to the value of the correlation coefficient.

ii. The second statistic used to measure discrimination was Kendall's Tau (B). The students were divided into three groups (TOP 27 %, MIDDLE 46 %, and LOWER 27 %), and the group rank of each student was matched with his number of choices for a given distractor type. Again, since the rank of the student is a negative function of the number of distractors chosen, the closer the value of Tau approaches - 1.0, the more effective the discriminating potency of the distractor.

iii. The final statistic used as a measure of discrimination was the contingency coefficient. This is an adjustment of Chi-square that measured the actual versus the expected frequencies for Bivariate Tables cross-tabulating student scores by the number of times a distractor was chosen. Since it measures nominal association rather than the relationships between ranks (Kendall's Tau) or scores (correlation coefficient), the contingency coefficient does not have a negative value for this data.

The following Table shows the values of these three statistics for each type of distractor. In each case the number in parentheses following the statistic is the rank of the distractor's ability to discriminate as indicated by that statistic.

TABLE #4 - DISCRIMINATION RANK FOR TOTAL POPULATION*

Distractor Type	Correlation Coefficient	Kendal's Tau B	Contingency Coefficient
Context	- 0.577 (1)	- 0.488 (1)	0.506 (1)
Cognate False	- 0.539 ( )	- 0.452 (2)	0.506 (2)
Affix	- 0.529 (3)	- 0.429 (3)	0.450 (3)
Antonym	- 0.463 (4)	- 0.353 (4)	0.388 (4)
Grapheme	- 0.455 (5)	- 0.352 (5)	0.380 (5)
Synonym False	- 0.441 (6)	- 0.338 (6)	0.379 (6)
Cloze	- 0.404 (7)	- 0.326 (7)	0.358 (7)
Arbitrary	- 0.389 (8)	- 0.286 (8)	0.329 (8)

* Figure and Ranking for each sub group may be found on Page #27, Table #7.

The distractor types are listed in an order reflecting their composite rating for the three statistics. While there is very close agreement among the three as to ranking for the total population, Kendall's Tau B is perhaps the easiest to defend in terms of both its appropriateness (the data is essentially a ranking when the students are split into groups of low-middle-high scores); and its consistency with the top 27% versus bottom 27% conventions used in item analysis.

Tables No. 5, 6 and 7 (pp 17, 26,27) provide summaries of the major data collected for the study. Table No. 5 provides Potency Data and Rank. Table No. 6 gives Discrimination Data and Rank, and Table No. 7 summarizes the Statistical Ranks of each Group.

INTERPRETATION OF DATA

An examination of the data shows good agreement between test samples for potency. All the students - good or bad, Saudi or Egyptian - generally agreed on the attractiveness of the specific distractor class. Such was not the case with discrimination. Some distractors proved more effective with the good students while the poor students found others more effective. The false synonym, for example, was very popular but not very effective at differentiation. It proved equally attractive to both good and bad students. And false cognates discriminated extremely well but were among the least popular choices. Only the poorest student found them appealing.

By itself the data is interesting but not very helpful. Some procedure must be established to provide the teacher with a useful index of choosing distractor class. Lacking evidence to provide guidance and for the sake of simplicity, we have combined rankings evenly. More sophisticated teachers might wish to give different weights to the potency and discrimination ranks. Based on the data presented here, one can ascertain visually those distractors which appear to be mose effective. They are, in order of rank, the following:

TABLE #5 :

Distractor	Potency	Discrimination
1. Context	2	1
2. Antonym	3	4
3. Synonym, false	1	6
4. Affix	5	3
5. Cognate, false	8	2
6. Grapheme	6	5
7. Cloze	4	7
8. Arbitrary	7	8

Samples chosen from the first four distractor groups should provide sufficient choices for any test question. One might make selections from all four or only one class and expect a better chance of having efficient distractors than if one were to choose items from words in the last four of the rank.

As a final step, our test was revised so that words from all the poor distractor groups were placed in Form A and words from all the good distractor classes were included in Form B. Both forms used 55 identical questions and answers retained from the original form and amended where necessary. The order of the questions differed but the answer choice position was identical for both forms. If our contention is correct that one group of distractors is more efficient than another, Form B should prove to be the better test.

The confirmation test was administered to 228 UPM students in May 1975. These students were almost identical to UPM Group II as to their general proficiency. The mean for Form A was 79.35. The mean for Form B was 74.1. The t-test for difference between the two means was significant at the 0.05 level (t=16.851). Both Forms were correlated with the students scores on three forms of the Michigan Test of Aural Comprehension, and the Michigan Test of General Proficiency (see Table #10, page 30). In all cases, the correlations were higher for Form B than for Form A. In essence, the B Form of the test proved better than the A Form. As the forms were identical with the exception of the distractors used, it seems safe to assume that the distractors helped to make the difference.

SUMMARY OF DISTRACTORS

In order to more easily examine constituent populations, Table #9 (page 29) has been prepared. This Table provides comparisons of ability with locality. The distractors are listed from left to right in order of the most to the least efficient for the total population. The Potency (P), Discrimination (D), and a composite Rank (R) is provided for each group .

9A - UPM I and UPM III are examples of similar populations but different proficiency.

9B - UPM I and AUC are examples of similar proficiency but disparate populations.

9C - UPM III and Dammam are two groups of similar proficiency and similar population.

It should be pointed out that less proficient students tend to guess more or to leave questions unanswered (a datum not provided for, unfortunately, in our program). This behavior leaves a less clear picture of their distractor preferences and the statistics tend to be of a lower order of magnitude. This is also born out by a closer examination of the Tau B cross tab print out (not given here) which shows a more scattered distribution of the subjects especially for the Dammam group.

What may we discover from this presentation? To return to our three original questions, we found that there does exist a definite hierarchy of distractor efficiency in terms of (1) potency and (2) discrimination. This was especially true for potency. All groups demonstrated a corresponding preference for distractor types. There was, however, an uneven agreement in discrimination. We also discovered that the most popular distractors were not necessarily the most discriminating. In other words, many test subjects were attracted, both good and bad, rather than just the bad as we might wish (9A and 9C). (3) We also found that subjects from different geographical locations (9B), which imply different dialects, may not have any major differences in terms of potency, discrimination or rank in their choices. Admittedly, this conclusion is weak in that only one alternate population was available and no comparable group was found for the lower population equivalent to either UPM III or Dammam.

An examination of Tables B and C shows that differences can be detected between groups of contrasting proficiency but similar populations. There appear to be major changes in distractor efficiency between the two populations which are most clearly identical in all but ability (UPM 1 and UPM III). There is less change with UPM III and Dammam, and almost none with UPM I and AUC.

The Context distractor ranked first with the upper group and fourth with the lower group. Its discrimination index was especially weak with the less capable though it remained quite popular.

Antonyms and Synonyms could be ranked equally for the total population but were reversed, strangely, in that Synonyms were more effective for UPM I while UPM III found Antonyms more attractive. Conversely, Antonyms failed to discriminate for the upper groups and Synonyms for the lower group though both distractors ranked equally for potency. It was an interesting and puzzling discovery but one for which we have no explanation.

An additional contrast for the less efficient distractors may be seen with the graphemic and cloze types. The former was ranked 4th with the more proficient students but 7th with the less able. The cloze ranked in reverse order - 6th and 3rd respectively. Why should good students have visual problems of discernment? It is equally perplexing.

The cloze distractors seemed to be strongest for the low and intermediate students. In the confirmation test (Table #10,

page 30) when ranked with the less efficient false cognate, grapheme variation and arbitrary distractors, the cloze type was an unambiguous first. One explanation for this showing is that this class of words proved in many cases to be identical or similar to words in the false synonym family.

The final three distractors agreed fairly evenly with all populations. The affix distractor was quite effective at all proficiency levels. It was a top discriminator but weak in popularity. However, this type may be misplaced in a vocabulary test and may be measuring different skills.

The false cognates were average in discrimination but near the bottom in potency. Only the very lowest students chose them. For that group only, they might be considered for inclusion in a multiple-choice vocabulary test.

The last distractor type - those chosen arbitrarily - demonstrated no redeeming features and should be withdrawn completely from consideration for any test of this nature.

CONCLUSION

In conclusion, we found that there were definite differences in the effectiveness of eight selected distractor types. We also found that the differences between populations of varied proficiency was greater than between groups from different geographical locations. As this was a descriptive study limited to native speakers of Arabic, one should not conclude that our results will be duplicated with other linguistically homogeneous populations. It is hoped that efforts of replication will be made with alternate language groups to provide additional data

to confirm or confound the results given here, but it seems safe to state that the conclusions of this study are indicative of a tendency for some types of distractors to be more efficient than others in certain kinds of test situations, and that teachers might construct better tests if they give more careful consideration to the distractors they wish to include.

FOOTNOTES

1. Different types of distractors discussed in the following:
   a) Clark, p. 77
   b) Harris, pp. 55-57
   c) Lado, pp. 191-194
   d) Valette, pp. 114 and 125
   e) Michigan Test Manual, p. 2

2. a) Testing Specialists Disagree. The parameters of acceptable potency for a question varies from 0.30 to 0.90. Some authorities insist on including questions of different levels of difficulty to spread the students. Others feel that keeping the difficulty level at the mid-point (70% if that is passing or 50% for a normal distribution) is ideal and students will spread themselves. The level of difficulty for a distractor is not given. But any alternate choice with a value less than 1% should be considered worthless. The upper value should vary depending on the number of distractors available for each question but certainly any value over 40% should be examined carefully.

   b) A question discriminates if it separates or differentiates high and low scoring students. The power of discrimination depends on the degree to which the wrong answers attract the misinformed student. Diederich (p. 2) states that

there should be at least a 10% (preferably 15%) difference between the high/low groups. Other statistics prefer different procedures but the principle remains.

3. Dr. Kenneth G. Pollock, Director, Data Processing Center, Dr. Fred Cromer, Director, Department of Statistical Services, and Mr. William McGinnis, Director, Educational Services who was especially helpful in the advice, design and programming of the project.

TEST POPULATION

1. AUC — American University of Cairo, tested Spring of 1974

   101 students, 18-40 years old.

2. UPM-I — University of Petroleum and Minerals, tested Spring of 1974

   247 students (Freshman and Sophmore) 18-21 years old

3. UPM-II — University of Petroleum and Minerals, tested Spring of 1974

   334 students (Orientation Year) 17-20 years old

4. UPM-III — University of Petroleum and Minerals, tested fall of 1974

   456 students (Orientation Year) 17-20 years old

5. UPM-IV — University of Petroleum and MInerals, tested Spring of 1975

   227 students (Orientation Year) 17-20 years old

6. Adult Education: Adult Education Group of students at University of Petroleum and MInerals, tested Spring of 1974

   25 students, 25-50 years old

7. Dammam — Dammam Gulf Secondary School, tested Spring of 1974

   40 students from the 10th grade, 15-16 years old

TABLE #6 ** POTENCY DATA FOR ALL GROUPS ** RANKINGS

Distractor	UPM I	UPM II	UPM III	AUC	Adult Education	Dammam	Total Population Rank
Context	+ 2.7	+2.8	+6.4	+1.9	+4.8	+4.6	
Rank	3	2	1	3	2	1	2
Cognate	- 6.1	-5.4	-4.2	-4.8	-5.1	-2.5	
Rank	8	8	7	8	6	6	8
Affix	- 2.5	-0.4	-0.8	-2.8	-0.6	0.0	
Rank	6	4	5	5	4	4	5
Antonym	+ 3.7	+2.3	+1.8	+3.3	+1.3	+1.9	
Rank	2	3	3	2	3	3	3
Grapheme	- 2.3	-1.7	-3.0	-4.5	-5.2	-4.6	
Rank	5	6	6	6	7	8	6
Synonym	+ 9.0	+5.2	+6.0	+10.2	+11.0	+3.9	
Rank	1	1	2	1	1	2	1
Cloze	+ 0.2	0.0	-0.2	+1.5	-0.8	-0.4	
Rank	4	5	4	4	5	5	4
Arbitrary	- 4.8	-2.8	-6.1	-4.7	-5.4	-3.0	
Rank	7	7	8	7	8	7	7

DATA							
Population	247	334	456	101	25	40	1203
Distractor Test Mean Score	80.1%	71.5%	55.5%	81.3%	65.3%	45.5%	67.7%
MTGP Mean Score	63	57	45	---	---	---	---

TABLE #7 ** DISCRIMINATION DATA FOR ALL GROUPS -- RANKINGS

Distractor	UPM I			UPM II			UPM III		
	Cor.	Tau B	Contin.	Cor.	Tau B	Cont.	Cor.	Tau B	Cont.
Context	-.714	-.500	.494	-.466	-.423	.447	-.274	-.128	.270
Rank	1	2	5	3	2	2	6	7	4
Cognate	-.656	-.473	.523	-.499	-.439	.460	-.325	-.221	.264
Rank	3	3	2	1	1	1	3	4	5
Affix	-.662	-.539	.620	-.482	-.374	.397	-.358	-.253	.297
Rank	2	1	1	2	3	3	2	1	1
Antonym	-.456	-.315	.339	-352	-.319	.353	-.362	-.230	.274
Rank	8	8	8	6	4	4	1	2	2
Grapheme	-.574	-.421	.431	-.436	-.307	.334	-.275	-.155	.203
Rank	5	5	7	4	5	5	5	6	8
Synonym	-.653	-.463	.455	-.295	-.212	.264	-.205	-.119	.213
Rank	4	4	6	7	7	7	8	8	6
Cloze	-.573	-.409	.516	-.266	-.210	.246	-.303	-.235	.271
Rank	6	7	4	8	8	8	4	3	3
Arbitrary	-.533	-.417	.521	-.415	-.263	.318	-.248	-.195	.208
Rank	7	6	3	5	6	6	7	5	7

Distractor	AUC			Adult Education			Dammam		
	Cor.	Tau B	Cont.	Cor.	Tau B	Cont.	Cor.	Tau B	Cont.
Context	-.517	-.368	.502	-.411	-.297	.372	-.497	-.276	.297
Rank	3	3	1	4	4	6	2	4	4
Cognate	-.585	-.475	.474	-.544	-.446	.452	-.466	-.412	.448
Rank	1	1	2	2	2	2	4	2	2
Affix	-.552	-.375	.456	-.634	-.542	.541	-.434	-.236	.279
Rank	2	2	3	1	1	1	5	6	5
Antonym	-.310	-.134	.259	-.348	-.206	.411	-.414	-.418	.465
Rank	7	8	8	6	7	5	7	1	1
Grapheme	-.491	-.342	.432	-.460	-.325	.425	-.471	-.214	.222
Rank	4	4	4	3	3	4	3	7	7
Synonym	-.322	-.295	.343	-.120	-.020	.354	-.428	-.242	.235
Rank	6	5	5	8	8	7	6	5	6
Cloze	-.331	-.238	.293	-.373	-.213	.352	-.394	-.143	.220
Rank	5	6	7	5	6	8	8	8	8
Arbitrary	-.284	-.187	.304	-.212	-.292	.439	-.605	-.342	.344
Rank	8	7	6	7	5	3	1	3	3

TABLE 8 - STATISTICAL SUMMARY BY RANK

		Totals and Averages	DAMMAM	UPM III	ADULT EDUCATION	UPM II	UPM I	AUC
A. Population figures for each sub-group in study		2203	40	456	25	334	247	101
B. Average score on Distractor Test in percent		67.7	45.5	55.5	65.3	71.5	80.1	81.3
C. Coordinate Evaluation Information:	Michigan Test of English Language Proficiency, Forms A - H			45		57	63	
D. Potency Rank	Synonym False	1	2	2	1	1	1	1
	Context	2	1	1	2	2	3	3
	Antonym	3	3	3	3	3	2	2
	Affix	4	4	5	4	4	6	5
	Cloze	5	5	4	5	5	4	4
	Grapheme Variation	6	8	6	7	6	5	6
	Cognate False	7	6	7	6	8	8	8
	Arbitrary	8	7	8	8	7	7	7
E. Discrimination Rank	Context	1	4	7	4	2	2	3
	Cognate False	2	2	4	2	1	3	1
	Affix	3	6	1	1	3	1	2
	Antonym	4	1	2	7	4	8	8
	Grapheme Variation	5	7	6	3	5	5	4
	Synonym False	6	5	8	8	7	4	5
	Cloze	7	8	3	6	8	7	6
	Arbitrary	8	3	5	5	6	6	7

TABLES # 9 - A, B, C

A. COMPARISON OF UPM I AND UPM III BY RANK AS EXAMPLES OF SIMILAR POPULATION BUT DISSIMILAR PROFICIENCY

	Context	Antonym	Synonym	Affix	Cognate	Graph-eme	Cloze	Arbi-trary
	(PDR)*	(PDR)	(PDR)	(PDR)	(PDR)	(PDR)	(PDR)	(PDR)
UPM I	3-2-1	2-8-4	1-4-1	6-1-3	8-3-6	5-5-4	4-7-6	7-6-8
UPM II	1-7-4	3-2-1	2-8-5	5-1-2	7-4-6	6-6-7	4-3-3	8-5-8

B. COMPARISON OF UPM I AND AUC BY RANK AS EXAMPLES OF SIMILAR PROFICIENCY BUT DISSIMILAR POPULATION

	Context	Antonym	Synonym	Affix	Cognate	Graph-eme	Cloze	Arbi-trary
	(PDR)	(PDR)	(PDR)	(PDR)	(PDR)	(PDR)	(PDR)	(PDR)
UPM I	3-2-1	2-8-7	1-4-1	6-1-3	8-3-6	5-5-4	4-7-8	7-6-8
AUC	3-3-1	2-8-5	1-5-1	5-2-3	8-1-4	6-4-5	4-6-5	7-7-8

C. COMPARISON OF UPM III AND DAMMAM BY RANK AS EXAMPLES OF SIMILAR PROFICIENCY AND SIMILAR POPULATION

	Context	Antonym	Synonym	Affix	Cognate	Graph-eme	Cloze	Arbi-trary
	(PDR)	(PDR)	(PDR)	(PDR)	(PDR)	(PDR)	(PDR)	(PDR)
UPM III	1-7-4	3-2-1	2-8-5	5-1-2	7-4-6	6-6-7	4-3-3	8-5-8
DAMMAM	1-4-2	3-1-1	2-5-3	4-6-5	6-2-4	8-7-8	5-8-7	7-3-5

* P = Potency Rank
D = Discrimination Rank
R = Rank Combined

TABLE # 10 *** CONFIRMATION TEST

DATA - UPM IV

	FORM "A" P*	D	Combined Rank	FORM "B" P	D	Combined Rank
1.	Cloze .0673	-0.734		Context .0632	-0.652	1
	Rank 1	1	1	Rank 1	1	
2.	Cognate .0448	-0.529		Synonym .0535	-0.584	
	Rank 2	3	2	Rank 2	2	2
3.	Grapheme .0385	-0.539		Affix .0457	-0.440	
	Rank 3	2	3	Rank 3	3	3
4.	Arbitrary .0260	-0.274		Antonym .0438	-0.440	
	Rank 4	4	4	Rank 4	4	4

Population = 118	Population = 109
Mean = 79.35 High: 98 Low : 28	Mean = 74.04 High: 99 Low : 13
Standard Deviation = 15.68	Standard Deviation = 17.03
Discrimination Mean = 0.4848 High: 0.799 Low -0.002	Discrimination Mean = 0.4353 High: 0.775 Low 0.0137

MICHIGAN TEST AC CORRELATIONS:				Combined Means
(Nov. 74)	1.	0.3332	0.6018	59.75
(Jan. 75)	2.	0.4240	0.6618	72.33
(May 75)	3.	0.3289	0.6565	78.67
MICHIGAN TEST G.P. CORRELATIONS:				
(May 75)		0.4740	0.5719	52.9
		$p \leq 0.001$	$p \leq 0.001$	

* Potency expressed in terms of total response average, e.g. cloze item chosen on average of 6% for all questions on Form "A".

BIBLIOGRAPHY

1. Bean, K.L., Construction of Educational and Personnel Tests, Greenwood Press, Westport, Conn. (1953).

2. Celce-Murcia, M., Kooshian Jr., G.B., and Gosak, A.J., "Goal: Good Multiple-Choice Language Test Items", ELT xxvii, 3, 257.

3. Clarke, M.A., "Arabic Distractors for English Vocabulary Tests", ELT xxvii, 1, 77.

4. Davies, Alan (ed.), Language Testing Symposium, Oxford University Press, London. (1968).

5. Diederich, Paul B., Short-Cut Statistics for Teacher-Made Tests, Educational Testing Service, Princeton, N.J. (1960).

6. Dixon, W.J. and , F.J., Introduction to Statistical Analysis, McGraw-Hill, New York City (1969).

7. Fowler, W.S., "Experimenting with Multiple Choice", ELT xxviii, 2, 145.

8. Garrett, H.E., Statistics in Psychology and Education (5th ed.)

9. Hamersma, R.J., Educational and Psychological Tests and Measures, Kendall-Hunt Publishing Co., Iowa (1972).

10. Harris, D.P., Testing English as a Second Language, McGraw-Hill, New York City (1969).

11. Hutchings, G., "Colourless Green Ideas; Multiple-Choice Vocabulary Tests", ELT xxv, 1, 68.

12. Katz, N., "Improving Classroom Test by Means of Item ANalysis", The Clearing House, January, 1961, pp. 265-69.

13. Lado, R., Language Testing (6th ed.), Longman, London (1970).

14. (a) Michigan Test of ENglish Language Proficiency Manual, University of MIchigan, (1962), Tests: Form A through H.

    (b) Michigan Test of Aural Comprehension. Forms 1, 2 and 3.

15. Naiman, Rosenfeld, & Zirkel, Understanding Statistics, McGraw Hill, New YOrk City (1972).

16. Sako, S., "Writing Proficiency and Achievement Tests", TESOL Journal 3, pp. 237-49, (1969).

17. Strevens, P., Papers in Language and Language Teaching, Oxford University Press, London (1965).

18. Thorndike, E.L. and Lorge, I., A Teachers' Word Book of 30,000 Words, Teachers College, Columbia University, New York City (1944).

19. Thorndike, R.L. (ed.), Educational Measurement (2nd ed.), American Council of Education, Washington, D.C. (1972).

20. Townsend, E.A., and Burke, P.J. Statistics for the Classroom Teacher, MacMillan, New York City (1963).

21. West, M. (ed.), A General Service List of English Words, Longman, London (1953).

22. Valette, R.M., Modern Language Testing, Harcourt Brace and World, Inc., New York City (1967).